Pipeline for the Analysis of ChIP-seq Data and New Motif Ranking Procedure

Thesis by

Haitham Ashoor

Submitted in Partial Fulfillment of the Requirements for the degree of

Master of Science

King Abdullah University of Science and Technology

Thuwal, Kingdom of Saudi Arabia

June, 2011

The thesis of Haitham Ashoor is approved by the examination committee.

Committee Chairperson: Vladimir Bajic

Committee Member: Mikhail Moshkov

Committee Member: Xiangliang Zhang

ABSTRACT

Pipeline for the Analysis of ChIP-seq Data and New Motif Ranking Procedure

Haitham Ashoor

This thesis presents a computational methodology for *ab-initio* identification of transcription factor binding sites based on ChIP-seq data. This method consists of three main steps, namely ChIP-seq data processing, motif discovery and models selection. A novel method for ranking the models of motifs identified in this process is proposed.

This method combines multiple factors in order to rank the provided candidate motifs. It combines the model coverage of the ChIP-seq fragments that contain motifs from which that model is built, the suitable background data made up of shuffled ChIP-seq fragments, and the p-value that resulted from evaluating the model on actual and background data.

Two ChIP-seq datasets retrieved from ENCODE project are used to evaluate and demonstrate the ability of the method to predict correct TFBSs with high precision. The first dataset relates to neuron-restrictive silencer factor, NRSF, while the second one corresponds to growth-associated binding protein, GABP. The pipeline system shows high precision prediction for both datasets, as in both cases the top ranked motif closely resembles the known motifs for the respective transcription factors.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor Professor Vladimir Bajic for his invaluable feedback, encouragement and guidance throughout this work. His enthusiasm for research and valuable guide made my study very enjoyable, exciting and ultimately fruitful with rich experience. I would also like to thank him for enabling me to work in an amazing research atmosphere.

I would like to extend my sincere thanks to Dr Boris Jankovic for his feedback and useful discussions throughout the work on this thesis.

I also would like to extend my thanks to Ehab Mousa for discussions in the biological aspects of this work.

I would also to express my deep gratitude to my parents for their continuous encouragement during this journey and their deep moral support at all times.

If it was not for the generous gift of the King Abdullah of Saudi Arabia, this work would not happen. I am truly grateful for the opportunity given to me.

TABLE OF CONTENTS

LIST OF ABBREVATIONS

bps	Bais pairs
ChIP	Chromatin Immunoprecipitation
DNase	Deoxyribonuclease
EM	Expectation Maximization
EMSA	Electrophoretic Mobility Shift Assay
ENCODE	Encyclopedia of DNA Elements
FDR	False Discovery Rate
GABP	Growth-associated Binding Protein
IC	Information Content
JBD	Joint Binding Deconvolution
KDE	Kernel Density Estimation
NRSF	Neuron-restrictive Silencer Factor
PCR	Polymerase Chain Reaction
PDF	Probability density function
PSSM	Position Specific Score Matrix

PWM	Position Weight Matrix
SELEX	Systematic Evolution of Ligands by Exponential Enrichment
TF	Transcription Factor
TFBS	Transcription Factor Binding Site

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter I

Introduction

I.1 Background

Genes play an essential role in any living cell. Their activity is necessary to sustain the vital living cell processes [25]. Genes activate differentially in response to different conditions or cellular demands. For example, humans have different types of cells. Different organs also have many cell types, like those in brain, liver, blood, etc. One of the differences between different cell types is that each is characterized by a specific set of genes active in that cell type. Other cell types require generally different set of genes to be active [27]. So, gene activities differ from one cell type to another and characterize different cell types. The level of activity of genes is known as gene expression. When a gene is active in the cell, it is said to be expressed, and if the gene is not active in the cell then it is said that the gene is not expressed. The process that controls gene expression is known as gene regulation [28].

Gene regulation process is mainly controlled by proteins called transcription factors (TFs) [16]. These proteins control when and how much genes express. In order to make such control on the gene, TFs have to bind to specific short sequence motifs on DNA [28]. These motifs are called transcription factor binding sites (TFBSs). Iden-

tifying these motifs can help in determining which TFs control which genes, which eventually help us to better understand gene regulation process.

TFBSs are short DNA sequences (motifs) with length in the range of 5 to 25 base pairs (bps) [40]. When several motifs are very similar to each other we can think of them as belonging to the same motif family. In this way TFBSs form many different motif families [41]. Individual TFs normally bind to the TFBSs of selected motif families [28].

Although there is a large number of human TFs, binding sites are known only for a small number of them [53, 2]. It is thus of great importance to identify from experiments the binding sites for TFs for which we still do not know TFBSs.

I.2 Motivation

Identification of TFBSs problem has been and remains an important field of research. Methods to solve this problem varied between experimental approaches and computational approaches. Recently, methods based on Chromatin Immunoprecipitation (ChIP) [34] were developed. One of these is the ChIP-sequencing (or ChIP-seq) method [52]. Generally, ChIP-seq is used to identify protein-DNA interaction, where the binding of TFs to TFBSs is one type of these interactions. ChIP-seq method is also used to identify histone modifications [50], etc.

Emerging ChIP-seq technology generates large amounts of data that contains TFBSs location for a given TF. While ChIP-seq data points to regions where TFs bind to DNA, it does not explicitly demarcate individual TFBSs. There are a number of systems/methods that identify TFBSs from ChIP-seq data [32, 36, 46].

In this study, we developed an integrated method to identify TFBS families from ChIP-seq determined binding regions. The method relies on the Dragon Motif Finder [7], a new parallelized version of the Dragon Motif Builder algorithm [17] that identi-

fies families of the short DNA sequences enriched in the DNA sequence sets as opposed to some background. A new (to the best of my knowledge) ranking algorithm that identifies the best TFBS motif family determined from ChIP-seq data is proposed. some of the advantages of the proposed pipeline system are that:

(a) it examines heuristically a very large space of motifs (motif families with motif length 5 to 20 bp) potentially enriched in the ChIP-seq data,

(b) it is very fast as it relies on the parallel version of the Dragon Motif Finder system.

The pipeline is successfully evaluated on the two ChIP-seq datasets, demonstrating its usefulness.

I.3 Problem Formulation

In order to define the problem of identifying TFBSs, we need to specify it in a more rigid mathematical setup. Let us first introduce several definitions relevant in the context of our problem. Before that let $\mathbb{R}_+$ stands for a set of positive real numbers.

Definition 1. *Let $N = \{A, C, G, T\}$ be an alphabet consisting of 4 characters A, C, G, and T.*

Definition 2. *A sequence is a string s of characters . A position j of a character c_j in the string s is determined by counting position of c_j from the start of the string s, with the first character in s having the position of 1. The length L of the string is determined by the position of the last character in s. A string s will be denoted by $(c_j) : c_j \in N, j = 1, .., L$, when necessary.*

Definition 3. *Let s be a sequence of length L governed by **Definition 2**. The reverse complement sequence r of s is a string r of characters obtained by: a) reversing the order of characters in s from the last position in s towards the first position in s, and*

b) replacing each of the characters A, C, G, T by T, G, C, A, respectively. Sequence s is called forward sequence relative to r.

Definition 4. *A motif m is a sequence that fits **Definition 2**.*

Definition 5. *A motif family M is a finite set of motifs having the same length.*

Definition 6. *A position weight matrix (PWM) $P = (p_{ij}), p_{ij} \in \mathbb{R}_+$, is a matrix obtained from a set of motifs that form a motif family M. The four rows correspond to A, C, G, and T characters, respectively. We denote $A = z_1$, $C = z_2$, $G = z_3$, and $T = z_4$. Each entry p_{ij} of the matrix P can be defined as:*

$$p_{ij} = \frac{f_{ij}}{\sum_{j=1}^{L} max_i(f_{ij})},$$

where f_{ij} is the frequency of the character c_j found at position j in all motifs in M.

Definition 7. *Let $P = (p_{ij})$ be a PWM with L columns. The Matching score, $Score : s \rightarrow ms \in \mathbb{R}_+$, is obtained by matching a sequence $s = (c_j)$ to P as follows:*

$$Score = \sum_{j=1}^{L} \sum_{i=1}^{4} p_{ij} \otimes c_j,$$

$$p_{ij} \otimes c_j = \begin{cases} p_{ij} & : c_j = z_i \\ 0 & : c_j \neq z_i \end{cases}$$

Definition 8. *Information content (IC_j) of a column j of PWM P=(p_{ij}) from **Definition 6** is defined as:*

$$IC_j = 2 - \sum_{i=1}^{4} p_{ij} log(p_{ij}).$$

Information content (IC) of P is defined as:

$$IC = \sum_{j=1}^{L} IC_j$$

Definition 9. *Let $M = \{m_j\}$ be a family of motifs m_j. M defines the threshold τ as:*

$$\tau = min_j(Score(m_j) : m_j \in M).$$

Definition 10. *Let S be a finite set of n sequences. Let n_t be the number of sequences in S that contain at least one motif from M. The coverage $C(M, S)$ of S by M is defined as n_t/n.*

Definition 11. *Let S_t and S_b be two given sets of sequences, with each of the sequences having the length greater or equal to L. If M is a set of motifs of length L, the significance of M is determined by a monotonously decreasing function*

$$F : p\text{-}value \rightarrow F(p\text{-}value) \in \mathbb{R}_+$$

where the p-value is determined based on the null hypothesis that $C(M, S_t) \leq C(M, S_b)$.

Now we can define the problems we deal with in determining TFBSs.

Problem 1. *Let $S = \{s_i\}$, where each s_i has length $L_i \geq L$, and L_i, L are positive integers. Let $\tau \in \mathbb{R}_+$ such that $0 < \tau \leq 1$. Find a family M of motifs of length L from sequences s_i and their reverse complements r_i, so that:*

$$max_M \left(IC(P(M)) : \, max\, C(M, S); \forall m_i \in M, \, Score(m_i) \, \geq \tau\right)$$

where P is a PWM defined by M.

Problem 2. *Let S_t and S_b be two given sets of sequences that fit **Definition 11**. Let M_j fit **Definition 11** and let $\{M_j\}$ be a finite set of such M_j. If F is a function from **Definition 11**, then rank M_j according to decreasing values determined by $C(M_j, S_t) \times F(p\text{-}value)$.*

The concepts defined in this section will take specific forms in our implementation of possible solutions of the above mentioned problems. For example, one family of TFBSs can be associated to one motif family M.

I.4 Thesis Organization

The remainder of this document is organized as follows: Chapter 2 discusses TFBS identification methods in general. Chapter 3 explains the computational analysis of ChIP-seq data. Chapter 4 describes all implementation details of the system, while Chapter 5 presents the results and discussions. Conclusions are given in Chapter 6.

Chapter II

Identification Methods for Transcription Factor Binding Sites

Since TFBSs are key components in the transcription regulation process, many methods were developed to identify such type of DNA motifs. These methods vary in their character: some of these are experimental, some are computational, while others are combination of experimental and computational (hybrid). This chapter contains short review of these three types of methods with some examples of each type.

II.1 Experimental Approaches

In a biological context TFBSs are short segments of DNA that bind regulatory proteins, TFs, and protein complexes they form. Many experimental methods were developed for identifying TFBSs. Examples of these methods include: Systematic Evolution of Ligands by Exponential Enrichment (SELEX) [10], DNase I Footprinting [14], X-ray crystallography [4], DNA-cellulose chromatography [43]. We will only make short reference to the first two methods.

SELEX method mimics the process of natural selection [10]. A large random group of DNA sequences of known length are generated to match a specific protein

targets, TFs in this case. Each of these sequences is surrounded by recognizable DNA ends (upstream end is called 5′ while downstream end is 3′) which serve as primers. Initially, it is attempted that all sequences bound to a selected target TF. Sequences that do not bind are excluded from further consideration. Sequences that bind with different affinity will be kept to the next stage. In the next stage the threshold of affinity is increased, leading to further elimination of candidate sequences. The same process is repeated until the all the remaining sequences show sufficiently strong binding affinity. The threshold of affinity will be increased each time. At the end the process all remaining sequences show strong binding affinity to the target.

DNase I Footprinting method was developed in [14]. This method relies on electrophoretic mobility shift assay (EMSA). The main idea in this method is to find the difference in the pattern generated by deoxyribonuclease (DNase) enzyme when the experiment is done on a binding region [31]. Usually, DNase breaks DNA into known pattern of fragments. But in the case when TF binds the TFBS, the dividing pattern will change, so that fragment can be identified as TFBS.

The main advantage for the experimental approach is the accurate results that it is providing. However, it has several disadvantages, such as: high cost, slow nature of the experimental results.

II.2 Computational Approaches

Computational methods are considered one of the important techniques to predict TFBSs. This section will introduce two key computational methods of these types. The first one uses genomic sequence motifs to build TFBS models and use such models to identify TFBSs, usually in the promoter region of genes. The other uses evolutionary information to identify TFBSs.

II.2.1 Position Weight Matrix

PWM [21], or Position Specific Score Matrix (PSSM), is considered the most popular way to model collections of TFBSs. Databases like TRANSFAC [53] and JASPAR [2] provide PWMs as models to predict TFBSs. As a model, PWM captures the frequencies of nucleotides in a group of DNA sequences at certain positions.

PWM model can be considered as a zero-order markov model. This means that a state at any position is independent from all other positions. The construction of PWM model is simple. Given a set of DNA sequences of length L, the occurrence of the possible states in that model is counted for each position. For DNA sequences there are four possible states which are A,C,G,T, Figure II.1 shows the process of construction a PWM. After that position weight matrix can be normalized, then each entry of the matrix represents the likelihood of the nucleotide at that position.

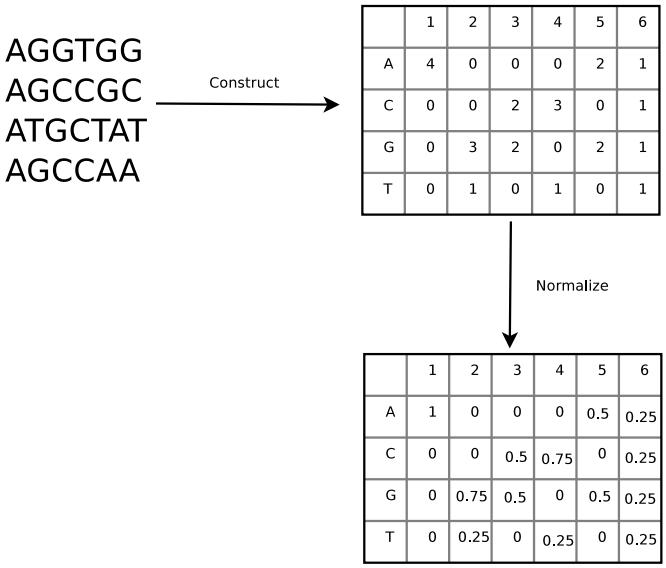
Given a sequence S of length L, we can calculate its matching score to the PWM as following: for each nucleotide in the sequence we sum the probability of observing that nucleotide at that position from the PWM. If the score is above certain threshold, the sequence is considered to match the matrix. The threshold usually is defined by the user of the algorithm.

Relations of TFBSs and PWMs are threefold:

(a) a PWM model can be built based on a group of real or candidate TFBSs; in this case the set of aligned TFBS motifs is used;

(b) one can use PWMs to identify candidate TFBSs in the DNA genomic sequence; target sequences are scanned with a sliding window, whose length corresponds to the PWM; the content of the window is matched to the PWM and a matching score for each window is calculated; DNA sequences contained in sliding windows with the score above certain threshold are considered as predicted TFBSs;

(c) one can try to identify PWM models from a set of longer sequences/fragments



	1	2	3	4	5	6
A	4	0	0	0	2	1
C	0	0	2	3	0	1
G	0	3	2	0	2	1
T	0	1	0	1	0	1

	1	2	3	4	5	6
A	1	0	0	0	0.5	0.25
C	0	0	0.5	0.75	0	0.25
G	0	0.75	0.5	0	0.5	0.25
T	0	0.25	0	0.25	0	0.25

Figure II.1: PWM construction example

that contain TFBSs; this can be implemented in many ways; one is based on the *ab initio* motif discovery algorithms which may use PWMs to build TFBS models, such as in methods based on use Gibbs sampling [12], Expectation Maximization (EM) [51], or greedy approach [22], etc

The most obvious trade-off in using PWMs for motif discovery is how to set the threshold properly to keep balance between the low false positive and the high true positive predictions. For example, maintaining high threshold will result in high quality predictions, but these predictions may miss many of the real motifs. On the other hand, reducing the threshold level will result in predicting more real binding sites, but in almost all cases it will result in a huge number of false positive sites.

II.2.2 Phylogenetic footprinting

Phylogenetic footprinting [15] identifies TFBSs from the conserved non-coding regulatory regions based on comparison from multiple species [40, 33]. The key argument behind this approach is that the conserved short sequences are likely to be important

(and thus they are conserved) and because they are in the non-coding regulatory regions they are likely to represent TFBSs. On the other hand, it is known that TFBSs can be conserved across some species [35].

A generic computational approach is to use global multiple alignment across the different species for specific regulatory regions, and then identify the conserved motifs across all species in these regions [35]. The critical step in the process of the phylogenetic footprinting is choosing the species to be used for the comparison [40]. The selection process should guarantee that evolutionary distance between any two species is sufficient to identify the conserved regions [33].

Many systems were developed to predict TFBSs using phylogenetic footprinting. These systems include: MicroFootPrinter [47] which is developed for prokaryotes. It uses BLAST program [49] to search for homology proteins for the desired gene. Another system is FootPrinter [35], which is based on dynamic programing paradigm to search for the conserved DNA segments.

In general, computational methods are faster and more cost effective than experimental methods. On the other hand it suffers from high false positive rate.

II.3 Hybrid Methods

Hybrid methods can considered as a trade-off between experimental methods' accuracy and the computational methods speed, also it can be cost effective if the correct technology is used in the correct case. Recent years witnessed emerging new methods for determining protein-DNA interaction based on ChIP technology [34]. ChIP process includes cross linking the DNA with the protein using formaldehyde, followed by DNA sonication, and after that the target proteins are precipitated using specific antibodies.

Genomics-based methods, such as microarrays and sequencing, are used to identify

TFBSs from DNA fragments resulting from ChIP. When ChIP method is used in combination with microarray hybridization, the resulting technology is called ChIP-on-chip or ChIP-chip [8]. Yet another method emerged with the development of high throughput sequencing technologies, where a sequencing process follows ChIP. This method is called ChIP-sequencing or ChIP-seq [52].

Both methods (ChIP-chip and ChIP-seq) produce a large volume of data that requires computational analysis to identify TFBSs in an accurate manner. Computational analysis includes: image processing for the microarray, sequence alignment, noise cancellation, signal peaks detection, and *ab initio* motif detection. The following sections will briefly introduce ChIP-chip and ChIP-seq methods.

II.3.1 ChIP-chip

As mentioned before, ChIP-chip consists of ChIP process followed by microarray hybridization. Computational analysis pipeline starts with image processing for microarray images, followed by the construction of signal peaks. Then, the abundance false signals can be reduced using False Discovery Rate (FDR) [9]. At this stage when the ChIP-chip peaks are identified the approach also identifies sequences (of length varying from tens of bps to several thousands bps) likely to contain TFBSs for the TF in question. To find these TFBSs any computational method for predicting TFBS can be used. This is usually done by identifying the most prominent motifs in these sequences.

Many computational methods were developed to process ChIP-chip data. rMAT [1] is an R package that uses empirical background distribution developed in MAT algorithm [54] to detect ChIP-chip signals also. Joint Binding Deconvolution (JBD) method [56] incorporates additional ChIP data to improve sensitivity and specificity. In MA2C [29], a normalization method was implemented based on the GC content of the microarray probes.

Nowadays, ChIP-chip technology is not widely used, because it is outperformed by the emerging of ChIP-seq technology [30]. A comparison between the two methods is discussed in section II.3.3.

II.3.2 ChIP-seq

Development of high throughput sequencing methods forms the basis of emerging ChIP-seq technology. ChIP-seq method consists of two stages. The first stage is ChIP, whilst in the second stage the associated sequences that are precipitated with the target TFs are sequenced using high throughput sequencing methods. Since ChIP-seq produces huge amount of data, a group of preprocessing and analysis steps has to be applied on this type of data. Chapter III discusses the computational analysis of this data in details.

II.3.3 ChIP-seq Vs ChIP-chip: advantages and disadvantages

ChIP-seq outperforms ChIP-chip because of several advantages [42]: First, ChIP-chip is considered to produce noisy data compared to ChIP-seq experiments, mainly because it suffers from cross hybridization between the probes. Also, different GC content between probes may lead to lower data quality. Second, ChIP-seq depends on tag count to measure the signal strength, while ChIP-chip depends on probe light intensity. This leads to the situation that ChIP-chip signal strength measurement is limited by probe saturation while there is no limit on tag counts. Third, the amount of ChIP data needed in ChIP-seq experiment is smaller than the one in ChIP-chip experiment. Fourth, the length of ChIP-seq output is shorter than ChIP-chip, where the range of peaks reported by ChIP-chip is between 50 to 300 bps, while the one for ChIP-seq is reported to be less than 50 bps. Finally, the coverage of ChIP-seq peaks

is greater than that of ChIP-chip because all repetitive regions in ChIP-chip will be excluded, but this is not always the case in ChIP-seq.

The primary disadvantage of ChIP-seq method is cost [42]. In general, the cost of ChIP-seq experiment is higher, but the technology can be cost effective when analyzing large amount of data as in the genome wide analysis. ChIP-chip is more cost effective when analyzing local areas from genome.

The last point to be made out is that all computational results have to be verified experimentally as the original purpose of the computational analysis is to identify candidates which most likely contain the TFBSs of the target TF.

Chapter III

Computational Analysis of ChIP-seq Data

Using the ChIP-seq technology we can perform genome-wide experiments regarding binding of TFs to DNA. These experiments produce high volume of data. This data requires further computational analysis to discern the TFBSs for the TF in question. Regardless the aims of ChIP-seq experiments the analysis of the experimental data is the same. Many computational approaches where developed to analyze this type of data and in this chapter we review some of these.

A general pipeline that analyzes ChIP-seq data consists of the following processes: tags alignment, profiling, peaks calling, and in our case *ab-initio* motif identification. There are many possible implementations for these procedures. In this section we will discuss the concepts behind these steps and different implementations as reported in the literature.

III.1 Tags Alignment

The process of tag alignment includes mapping of ChIP-seq fragments to the reference genome. The alignment of ChIP-seq DNA fragments (tags) is an example of alignment

of small length sequences to long references. This process is computationally intensive and may take long time. Making correct heuristics and reference indexing will increase the speed of alignment multiple times. There are many examples of alignment tools that are specialized for this task such as BLAT [55] and BWA [24].

The main issues in the alignment process regarding ChIP-seq data are that a tag can be mapped into multiple locations of the genome and that multiple tags can be mapped into the same locations of the genome. The simplest solution of the first problem is to ignore the tags mapped to multiple locations in the genome, and for the second problem to take one instance of multiple tags aligned to the same location. More complex solutions can also be implemented. For example, one can use ClustalW [38]. This is software for multiple sequence alignment that uses probabilistic models to give a score to each location that a tag is mapped into and the location with the highest score can be selected as the mapped location. The multiple tags mapped to the same location are most likely based on polymerase chain reaction (PCR). A model based method was implemented in [57] to remove duplicates for the same location.

III.2 Profiling

The main purpose of profiling is to represent the aligned tags in a form that facilitate the process of determining regions enriched by mapped ChIP-seq data from others. Another purpose of profiling is to smooth the signal generated by the aligned tags [48]. Many techniques are used to build profile from the aligned data. Examples of these are: window scan [23, 57, 11], tags aggregation [20, 6], and kernel density estimation [5, 3].

In the window scan method, which is considered the simplest, a window with fixed width will slide along the whole genome and replace the tag count at each window position with the sum of ChIP-seq tags included in the window centered at

that position [48]. Many variants of this method were implemented with ChIP-seq analysis systems. For example, SiSSRs [44] implemented the simplest version of this method as described above; on the other hand, CisGenome [23] is using strand specific profiling where each strand has its own profile and then the two profiles are merged in one; MACS [57] shifts tags before window scanning, while SICER [11] allows gaps during the scanning.

Tags aggregation method combines all overlapped ChIP-seq data into one signal. GLITR [20] and FindPeaks [6] are examples of systems that implement this method. Another variant of tag aggregation is used in XSET [19] that extends tags to the expected fragment length and then aggregates them. An example of systems that are using this method is PeakSeq [26].

KDE is a non-parametric method that estimates the probability density function (PDF) and it is also used as a signal smoother. F-Seq [5] and QuEST [3] use different forms of KDE with Gaussian kernel.

In the case of strand specific profiling, a shift operation is performed after constructing strands profiles. The main purpose of shifting is to merge strands profile into one profile in the case of strand specific profiling. Figure III.1 shows positive strand profile before shifting, negative strand profile before shifting, and the merged profile after shifting both distributions towards the center. In that figure the blue distribution represents the positive strand, the green distribution represents the negative strand, and the final profile is represented by the red distribution.

Many approaches were used to estimate the shift for the strands. In CisGenome [23] the peak shift is estimated based only on the high quality peaks; in QuEST [3] the shift is estimated as the distance that maximizes the cross-correlation between the peaks from the positive strand and negative strand. Another approach is to make the shift as an input parameter to the system such as in SICER [11] and then computationally determine the best one. The main problem of the shift estimation

process is that if the shift is underestimated or overestimated some real peaks could be missed [48].

In contrast, in this study shifting was not implementing due to the following reasons:

(a) estimation of shift may lead to wrong estimate,

(b) it is possible to derive ChIP-seq fragments corresponding to peaks without performing shifting.

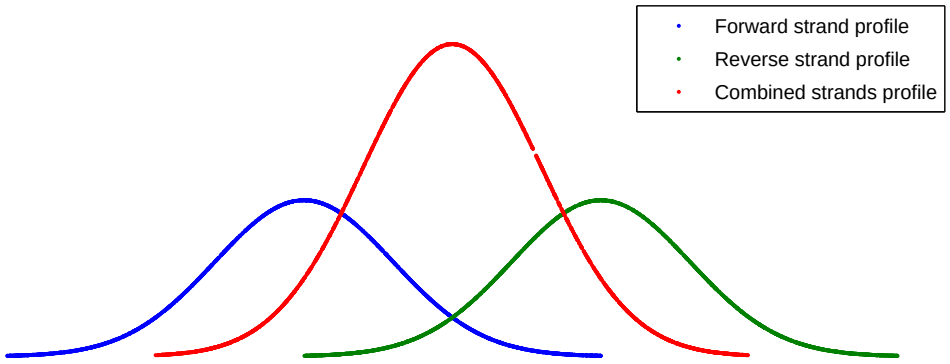


Figure III.1: Peak shift and profile merging

III.3 Peak Calling

This procedure makes selection of peaks and in this manner influences selection of real peaks. The aim is to select most of the real peaks at the expense of the false positive peaks. Real peaks are corresponding to ends for the binding regions for both strands [57], while in most cases false positive peaks are corresponding to a contamination process that may happen during the experiment. Usually a ranking procedure takes place to identify best peaks among all other peaks. Peaks are ranked based on p-values as in MACS [57], and SiSSRS [44]. Other systems are using q-value as the

ranking score such as PeasSeq [26] and QuEST [3]. The other way to rank peaks is just to rank it by the maximum number of overlapped tags as in CisGenome [23] and F-Seq [5].

A preprocessing step comes before the actual peaks calling. This procedure is associated with FDR calculation [9]. The process varies between systems in terms of implementation. It also varies within the same system depending on the ChIP-seq experiment type depending on whether it is one sample experiment or two samples experiment. In a two samples experiment the control data is present. The control data in most of the cases is generated by doing the experiment without including any antibody targets, which are used to precipitate the binding regions during ChIP process.

The most common form of the FDR in the case of two samples experiment is computed as following: the number of tags in a ChIP-seq peaks, and the number of control tags that are mapped to the same peak region are calculated. Then FDR is calculated as the ratio between the number of control tags and the number of ChIP-seq tags in that peak. Systems like GLITR [20], MACS [57], and QuEST [3], are using this method. CisGenome [23] is using conditional Binomial distribution, while SICER [11] is using Poisson distribution p-values as FDR.

In the case of the one sample experiment many methods were used. The most popular two are the Poisson distribution background as in PeakSeq [26], and Monte Carlo simulation as in FindPeaks [6]. On the other hand, some of the systems do not use FDR in their algorithms, such as is the case with F-seq [5] where FDR is not used neither for one experiment nor for two experiments data.

III.4 *Ab-initio* Motif Identification

Motif identification is the process of selecting over-represented motifs (motif families) in a group of sequences. Finding over-represented motifs is considered as a combinatorial problem [39]. This means that we need to enumerate all possibilities to find the over representative motifs in a group of sequences. By this fact we can see that the complexity of this combinatorial approach is growing in an exponential rate. Also, motif finding problem can be classified as an NP-complete problem.

Many methods were develop to approximate this problem, an example of algorithms that approximate these motifs are: Gibbs Sampling [12], EM [51], and greedy approach[22].

Gibbs Sampling approach uses PWM as the model. It selects one sequence at the time to improve its PWM model. Initially the Gibbs Sampling algorithm selects a random motif with length L from each input sequence, then it constructs PWM model out of those motifs. It also constructs a background model by removing the selected motifs from sequences, then it constructs the background PWM as a sum of frequencies of each nucleotide in all positions.

After constructing the initial model, the algorithm samples one sequence in each iteration. At every iteration, the algorithm will remove the selected motif corresponding to that sequence from the list of selected motifs generated from the previous iteration, then it will reconstruct its PWM. The sequence will be evaluated by the new model using a sliding window and a weight is calculated for each window. A new motif will be randomly selected from that sequence to add to the PWM based on the calculated weights. The algorithm stops when there is no improvement on the constructed model.

EM also uses PWM to construct its model. It consists of two stages, the expectation stage where the model and its parameters are recalculated from the previous iteration, and the maximization stage where the model and its parameters are refined

until some convergence criteria are met.

Initially, the algorithm picks random motifs from each input sequence and build its PWM, it also initialize its parameters randomly. In the expectation stage the constructed PWM is evaluated through all sequences using a sliding window. In the maximization stage motifs with the highest scores in each sequence are used to reconstruct the PWM model and the algorithm also refines its parameters in this stage. The algorithm will stop when there is no improvement of the model.

In the greedy approach, the PWM model is refined in a greedy fashion which means that the algorithm will apply local optimization on each step to find the model. Initially, the algorithm starts with one sequence, then it starts adding one sequence in each iteration. The new added sequence will satisfy the condition of maximizing the information content at that iteration.

Chapter IV

Pipeline Design and Implementation

This chapter will explore the details of the design and implementation of the system for identification of TFBSs from the ChIP-seq peak fragments. The system is implemented as a pipeline where each processing stage depends on the output of the previous processing stage. Figure IV.1 shows the block diagram of the system that contains five subsystems starting from strands profiling and ending with reporting the best models that represent the model of presumed TFBSs of the targeted TF. Each of these subsystems will be discussed in a separate section.

IV.1 Signal Profiling

The purpose of the signal profiling process is to align input ChIP-seq fragments in a form that will allow easier detection of potential motifs. The profiling strategy used in this study is tags aggregation followed by window scan for each group of the aggregated data in order to smooth the overall signal. The reason of using this method is the time-efficiency compared to the window scan and KDE methods, since in each of these latter two methods one needs to scan the whole genome. In the

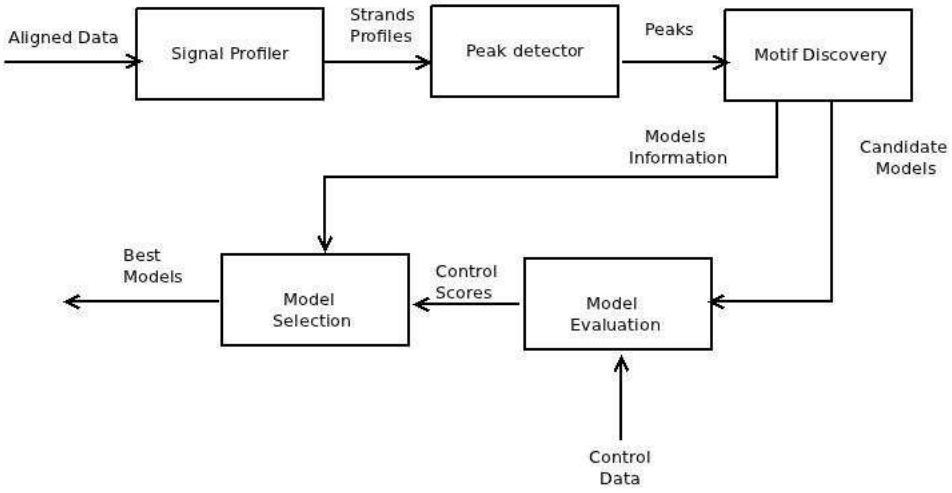


Figure IV.1: System block diagram

case of tags aggregation, however, the algorithm only processes the alignment files. Algorithm 1 is described by the pseudo code of the profiling algorithm

Algorithm 1 : **Profiler Algorithm**

Require: tag_{1-n}
  $i \leftarrow 1$
  append($peak_i$, tag_1)
  for all tag_j in strand where $j > 1$ **do**
    if $start(tag_j) \leq end(tag_j - 1)$ **then**
      append($peak_i$, tag_j)
    else
      for all $wnidow_j$ in $peak_i$ **do**
        $density_{ij} = \sum_{j}^{j+win\_length} nucleotides\_within\_the\_window$
      end for
      $i \leftarrow i + 1$
      append($peaks_i$, tag_j)
    end if
  end for
  return peaks,density

The algorithm will check all the overlapped ChIP-seq tags. Then, it will construct the signal using the window scanning method with window length equal to approximately the third of the fragment length. After constructing the signal, the average density of this signal is calculated as the average value of densities among all windows

in the signal as in equation IV.1. This algorithm is applied on each strand for each chromosome.

$$Avg_{density} = \frac{\sum_{j=1}^{n} density_j}{n} \tag{IV.1}$$

Figure IV.2 illustrates the profiling process. At the beginning a group of aligned ChIP-seq tags will be aggregated into groups. After that a sliding window, (of length 2 in this example), will scan through that group. It will assign each start index of the window with the number of nucleotides in that window. At the end, the signal will be constructed as at the right most of Figure IV.2. Then for each constructed signal the average density will be constructed according to equation IV.1. The average density will be used as the scoring criteria of the signal. Since the most dense regions in ChIP-seq represent regions with high probability of containing the binding site, average density will give a good indication for these regions. For example, if there is a group of aggregated tags that are loosely overlapped then the average density will be small; however if the group has higher overlap then the average density will be higher.

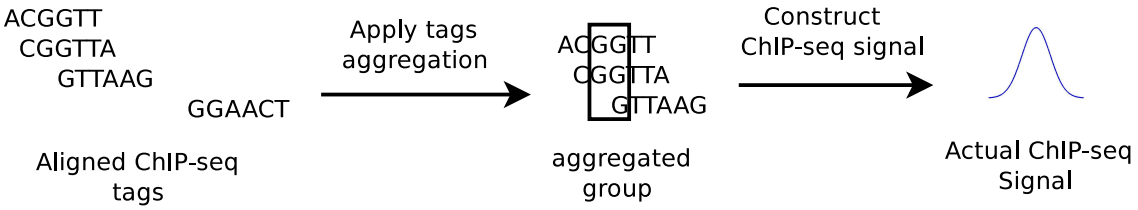


Figure IV.2: Profiling process

IV.2 Peaks Calling

In this stage the most enriched peaks will be predicted based on specified criteria. Peak detection module call peaks for each chromosome separately. The criteria that

is used to call peaks is based on two conditions: the first one is that the peak shows a bimodal distribution (this means that the peaks should be present in the forward strand and the reverse strand as well), this is because ChIP-seq tags represents fragments from both strands [57]; the second criteria is that the enrichment of the peak should exceed selected threshold for peaks in both strands.

Figure IV.3, shows all the cases that the peaks calling algorithm will deal with when at least one of the peaks is over the threshold. The first case is when the peak is present in one strand only. In this case the peak will be rejected because it does not satisfy the bimodal distribution condition. In second the case, the peak has the bimodal distribution property, but the peak in the reverse strand has a density smaller than the selected threshold, so the peak will be rejected. Finally the last peak will be called since it satisfies both conditions. After calling the peaks, a score of each peak is calculated as the weighted average of both strands average densities. After that, the top N peaks will be selected for further processing.The called peaks represent the longer DNA fragments that ideally contain TFBSs (motifs) that we look for.

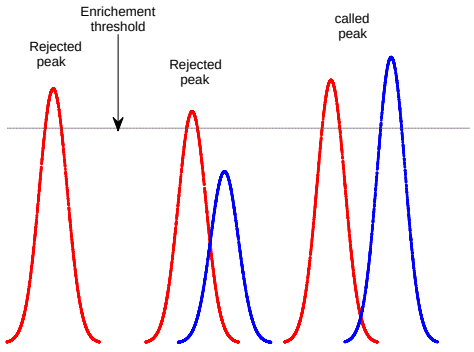


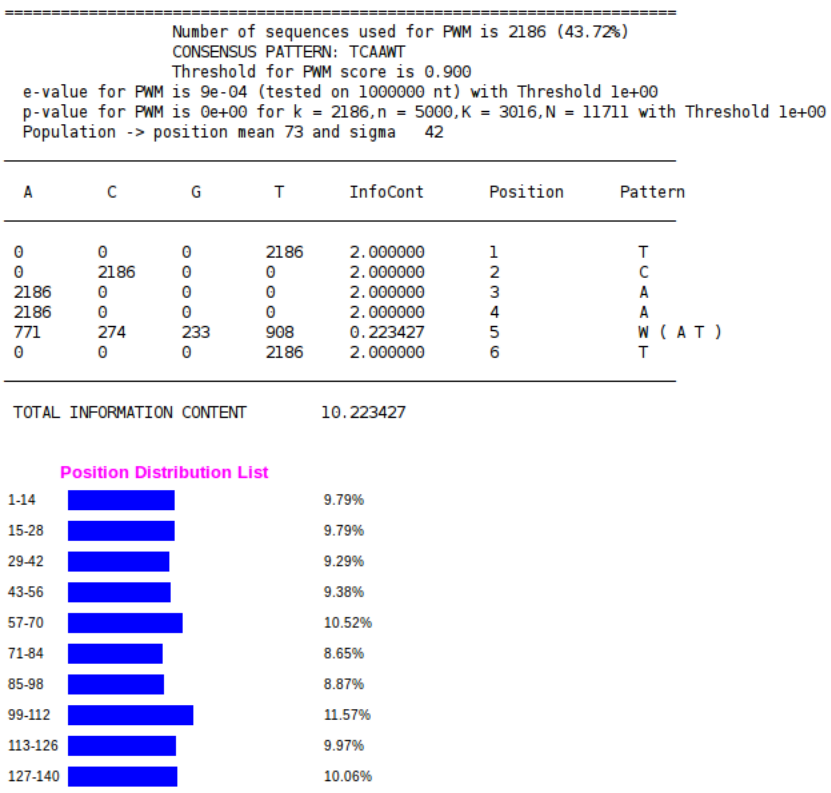
Figure IV.3: Peak calling criteria

IV.3 Motif Discovery

Motif Discovery process represents a core process in the system. In this study Dragon Motif Finer [7] was used to derive the candidate TFBSs from the selected peaks.

Dragon Motif Finder algorithm is a parallelized version of the Dragon Motif Builder [17] but otherwise follows the same logic. The algorithm is based on the EM concept. The algorithm starts with a collection of sets of randomly randomly compiled motifs. Each set of motifs is derived form a randomly selected motifs, one from each of the fragments. The PWM model is derived for each of the motif set. For each of the initial PWM models, the IC is calculated. The model with the highest IC is chosen; then an improvement of the motif family captured by the model is made. All DNA fragments will be scanned along both strands and motifs that have the highest score above the selected threshold in each fragment will be used to construct the new PWM of that model. The PWM threshold is provided as input argument for the algorithm. This process goes iteratively until no change in the model can be obtained. The algorithm continues selecting models based on IC (and optionally other parameters such as probability of finding motifs in the background) until the desired number of motif families is identified. A snapshot of one of the reports that Dragon Motif Finder produces is shown in Figure IV.4.

The Dragon Motif Finder algorithm will search for all motif families of specified motif lengths and will model them by PWMs. The other factor in the search process is the threshold for PWM score, as discussed in section II.2.1. Here is a trade off in the choice PWM threshold because high threshold will provide high IC models, but low coverage in the ChIP-seq data. On the other hand, the choice of low threshold will result in low IC models, thus resulting in the selection of the candidate TFBS motifs that show great mutual variability, although the coverage will be very high.



```
=====================================================================
                Number of sequences used for PWM is 2186 (43.72%)
                CONSENSUS PATTERN: TCAAWT
                Threshold for PWM score is 0.900
 e-value for PWM is 9e-04 (tested on 1000000 nt) with Threshold 1e+00
 p-value for PWM is 0e+00 for k = 2186,n = 5000,K = 3016,N = 11711 with Threshold 1e+00
 Population -> position mean 73 and sigma   42
```

A	C	G	T	InfoCont	Position	Pattern
0	0	0	2186	2.000000	1	T
0	2186	0	0	2.000000	2	C
2186	0	0	0	2.000000	3	A
2186	0	0	0	2.000000	4	A
771	274	233	908	0.223427	5	W (A T)
0	0	0	2186	2.000000	6	T

TOTAL INFORMATION CONTENT 10.223427

Figure IV.4: Dragon Motif Finder sample output

IV.4 Model Evaluation

This stage relates to the assessment of the model performance using control and test sequences and the candidate models obtained from the motif discovery process. One preprocessing stage, the PWM normalization, is required before the model assessment. The aim of normalization is to produce a PWM that will generate matching scores in the predefined range from 0 to 1. Here, the higher score corresponds to the better matching of motif to the model.The normalization process can be described as in equations IV.2 and IV.3.

$$factor = \sum_{i} max_j(matrix_i) \tag{IV.2}$$

and

$$matrix_{ij} = \frac{matrix_{ij}}{factor} \tag{IV.3}$$

After the normalization process the possible score range for matching the DNA sequences with the PWM model is between 0 and 1, the 1 representing the highest score (this corresponds to the best matched sequence). The model matching algorithm is summarized as in Algorithm 2. The algorithm receives as input the PWM model and the desired threshold for the score, as well as the sequences to be assessed i.e. scanned by the model. Then each sequence is scanned using a sliding window that corresponds to the matrix. Each nucleotide in each of the windows obtained sliding along the sequences is matched to the element of the matrix on the corresponding position and the matched outcome is contributing to the overall score. Then, if the matching score of the sequence in the window exceeds the threshold the sequence is flagged as covered by that model (i.e. there is a motif from the motif family described by PWM that can be found in the analyzed sequence).

Algorithm 2 : **Matrix Matching**

Require: model,sequences
  for all $sequence_i$ in $sequences$ **do**
    $r\_sequence_i = reverse\_complement(sequence_i)$
    for all $sliding\_window_j$ in $sequence_i$ **and** $r\_sequence_i$ **do**
      $score = \sum model_{wl} \otimes n_l$
      if $score > threshold$ **then**
        $hits_i = 1$
      end if
    end for
  end for
  return $hits$

IV.5 Model Selection

Model selection is the final stage in the developed pipeline system. It incorporates information from the model evaluation stage and the models details to set up ranking

criteria that will take all system variables into consideration accompanied with applying the appropriate trade-offs. The model selection process consists of two sub-stages: the p-value filtering, and the models ranking. This model selection/ranking process of motif models in ChIP-seq data analysis is novel to the best of my knowledge.

Figure IV.5 shows the two sub-stages of the system accompanied with its inputs and outputs. Using both sub-stages ensures selecting models that show higher enrichment over control data, and higher coverage in ChIP-seq peaks data. The next subsections will present the details of each of the sub-stages.

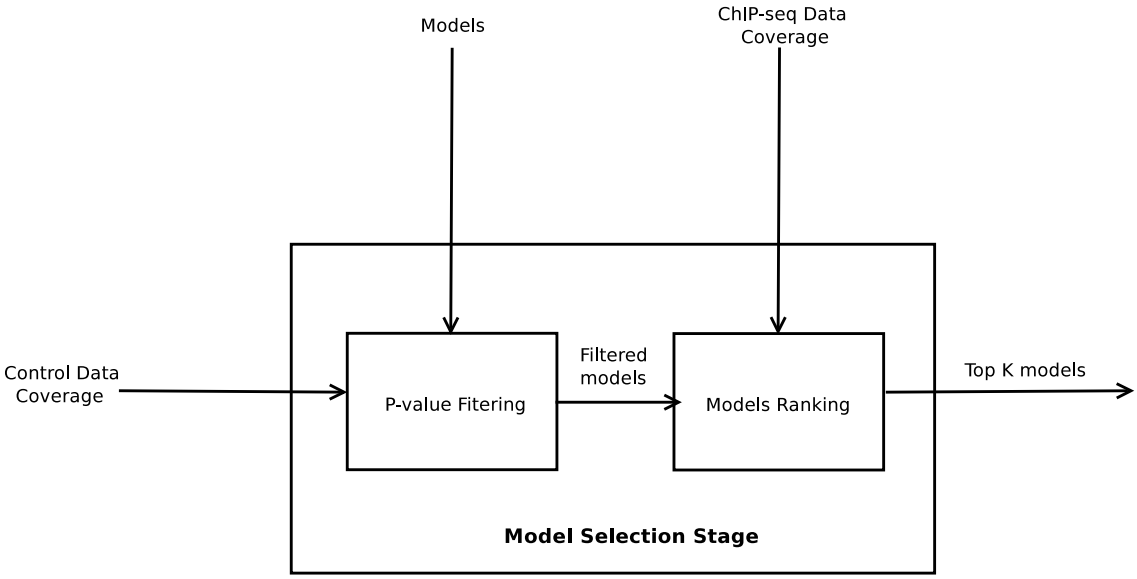


Figure IV.5: Model Selection Stage

IV.5.1 P-value Filtering

The p-value is the likelihood of observing a test statistic (under the null hypothesis) as large as the one calculated from observations. Here, in our context, the null hypothesis is formulated as follows: the coverage of a certain model when applied to the actual data and when applied to the background data is the same. The coverage can be defined as the number of the peak fragments predicted by the model to contain the

respective TFBS, and divided by the total number of peak fragments. Hence, p-value represents the probability that this hypothesis is true.

In literature many statistical tests exist to compute p-values, such as binomial test, z-test, student test, Fishers exact test, chi-square test, etc. In this study Fishers exact test is used to compute p-values. In order to calculate p-values, a contingency table is constructed. Figure IV.6 shows the contingency table for the case of model selection.

Number of ChIP-seq fragments are covered by the model (a)	Number of control fragments are covered by the model (b)
Number of ChIP-seq fragments are not covered by the model (c)	Number of control fragments are not covered by the model (d)

Figure IV.6: Contingency table for p_value filtering

We can see that the data in each population is split into two categories: fragment that are covered by a model, and the ones that are not covered. Each of these categories is labeled from a to d. Then, p-value can be calculated from the hypergeometric distribution as shown in equation IV.4 [18].

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{IV.4}$$

where

$$n = a + b + c + d \tag{IV.5}$$

Since multiple tests were conducted, the calculated p-value should be corrected

for multiplicity testing. We used a simple Bonferroni correction method [45]. Then, the models with the corrected p-value less than the user specified threshold will be selected to enter the next stage where the models will be ranked.

IV.5.2 Models Ranking

In the model ranking process, the filtered models that show enrichment in the target data over the control samples will be processed further. The model ranking will take into consideration two variables to compute the score of each model. These variables are the model sensitivity on ChIP-seq data (sensitivity), and the models p-value. In this study, we introduce the following method for ranking models as described in Algorithm 3.

Algorithm 3 : **Models ranking**

Require: models, p-values,p-vlaue-threshold
  for all $model\_i$ in $models$ **do**
    if $p\_value_i \neq 0$ **then**
      $Score_i = sensitivity_i \times \log_{10}(10 + \log_{10}(\frac{p\_value\_threshold}{p\_value_i}))$
    else
      $append(zero\_models, model_i)$
    end if
  end for
  $max\_score = max(Score)$
  for all $model_j$ in $zero\_models$ **do**
    $Z\_score_j = max\_socre + sensitivity_i$
  end for
  $all\_scores = combine(Score, Z\_score)$
  $sort(all\_scores)$
  return all_scores

As follows from Algorithm 3 the ranking process combines multiple factors to enhance the model ranking process. The first term in the equation represents sensitivity over ChIP-seq data. This sensitivity can be obtained as the average of the model coverage on the training data and the testing data. The last term represents the scaled p-value. The main reason of p-value scaling is to transform the original p-values into

a range that is comparable to other factors, since the p-value is frequently too small compared to the sensitivity.

Also, the algorithm handles the case when p-value is equal to zero. Simply, the algorithm ensures that models with p-value equal to zero are ranked at the top (the top being the best). This is because p-value of zero usually comes from high coverage of the real data and very low coverage (that may reach zero coverage) of the background data. Since multiple models can have p-value zero, their mutual ranking is based on the coverage.

For the other cases, the way how algorithm determined the model preference is to compute the score for all models where p-value is not zero. Then it will add the coverage to the maximum score achieved in the computed scores. In the application on two ChIP-seq data, the algorithm places on the top the motifs we expect to find.

Chapter V

Computational Experiments, Results, and Discussion

This chapter will describe the computational framework that has been set to test the pipeline discussed in Chapter IV, After that, the results are presented and discussed.

V.1 Computational Experiments

This section will describe the experiments used to evaluate the pipeline. First, data used in these experiments will be described. Then, all variants of experiments will be described as well.

V.1.1 Datasets

Datasets used in this study are for human TFBSs for which their TFBSs have been partly known and thus they have their TFBS models in the form of PWMs. Because of this, it is possible to assess if the models identified by our pipeline are good or not. Data was retrieved from the Encyclopedia of DNA Elements (ENCODE) [13] that can be found at a public repository of ChIP-seq data (http://genome.ucsc.edu/ENCODE/downloads.htn

Datasets include data for two TFs targets. The first data set corresponds to TF known as neuron-restrictive silencer factor (NRSF) and the ChIP-seq experiments were made using BE2_C cell line (Human neuroblastoma cells). The second dataset corresponds to TF known as growth-associated binding protein (GABP) with experiments done using Gm12878 cell line (lymphoblastiod). All dataset files are ChIP-seq fragments aligned to the human genome. They are given in the BED file format.

V.1.2 Experimentation's Flow

All experiments start with building the profile for each dataset. Then peaks of each datasets are called. After calling ChIP-seq peaks, the actual sequences corresponding to the peaks are extracted from the reference human genome (hg19). Then, the sequences (regions) corresponding to the top 1500 peaks (the most enriched ones) from each dataset are extracted. The 1500 sequences are divided into two sets of equal sizes. One set is referred to as the training set, while the other as the testing set. Usually, the splitting process is performed randomly, but in this study the split is made as follows: from the 1500 top ranked sequences, every second is selected to be in the test set. The remaining ones made the training set. The split was made in that way to keep the quality content of sequences balanced between the training and the testing sets.

After that, the training set is used to build the TFBSs models using Dragon Motif Finder tool. An exhaustive search was made by generating 100 models for each motif length from 5-20 bp, and for each PWM threshold value from 0.7-0.9 using the step of 0.05. This leads to a total of 8000 models generated for each TF target dataset. The next stage is the model evaluation on the background data and the testing data. Here, the p-values are calculated from the coverage values reported by the model, and the coverage values from applying the model to the background data. Test coverage is calculated by evaluating models on the testing data. Two backgrounds variants

were used to determine the p-values. The first one is the control data for the cell line for which the ChIP-seq experiment was made. The same number of control peaks as the number of the training peaks is selected randomly from the control data.

The second variant of the background data is a shuffled version of the training data itself. The shuffling process was performed using a program called uShuffle [37]. This program performs random shuffling of the sequence while keeping a k-mer content conserved. This program performs random shuffling of the sequence while keeping a k-mer content conserved. The shuffling that was performed kept 1-mer (mononucleotide) content conserved in the training sequences.

After that, for each background type the model is selected as discussed in Section IV.5. So, at the end, two rankings were produced, one based on the cell line control data, and the other based on shuffled training sequenced data.

V.2 Results and Discussion

This section will show the results obtained by applying the pipeline system to the NRSF and GABP ChIP-seq datasets. Tables V.1 - V.4 resent the top 5 TFBSs found by the method described in Chapter IV. Results are reported for both shuffled background and the cell line control background.

Table V.1 shows the top ranked TFBSs for NRSF TF based on shuffled background. The first motif in the table shows high similarity to the reported TFBS from TRANSFAC. Figure V.1 shows both binding sites in a form of sequence logo (where the x-axis represents the position in the motif and y-axis represents the IC of that position). One can see that the sequence reported by our system from positions 1 to 15 is very similar to the sequence reported by TRANSFAC from position 3 to 17. There is no possibility to correctly compare the IC for the two collections of binding sites represented in Figure V.1 because TRANSFAC binding site model is built from

only 21 TFBS sequences likely to introduce a strong bias at some positions, while our model is build from 627 sequences. It can be observed that the length of the binding site identified by our system from the ChIP-seq data is 15, while the binding site reported by TRANSFAC from experimental resources is 21.

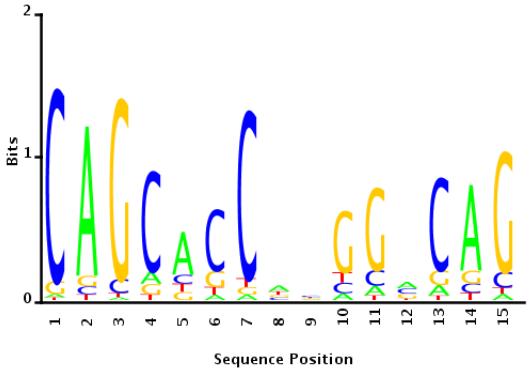
Table V.1: Top 5 ranked NRSF TFBSs based on shuffled background

Predicted model consensus	Score	P-value	Train coverage	Test coverage
CAGCACCAYGGACAG	1.35	3.74151e-39	0.837	0.806
CTGCTCT	1.33	7.3169e-20	0.968	0.956
AGGCAGGGG	1.32	1.41629e-23	0.921	0.916
TGCTGA	1.29	6.34717e-19	0.926	0.973
CTCTGCCT	1.24	9.52286e-26	0.763	0.929

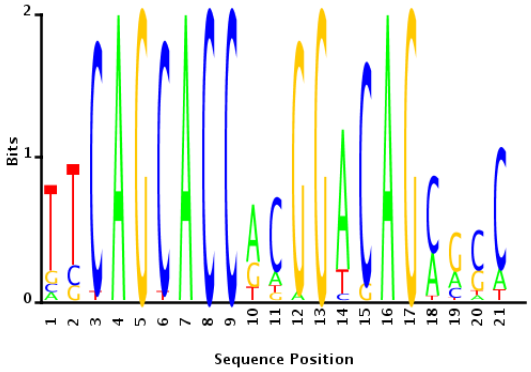
Another conclusion can be made from Table V.1. The motifs that follow the top motif show high coverage with small p-values as well. One explanation is that there are multiple TFBSs that are close to NRSF binding sites. This suggests that possibly other TFs that bind such TFBSs are required for the proper activity of NRSF. This is hypothesis and it requires further detailed analysis. Another explanation could be that NRSF could actually bind these other TFBSs in which case these would be the new models of NRSF binding sites. This is also a hypothesis that requires detailed experimental validation.

In Table V.2 the binding sites for NRSF TF are reported, using cell line control data as a background. One can see that the consensus motif for the first (top) model differs from the one reported in TRANSFAC. However, partial similarities between other motifs reported using our approach and motifs reported using shuffled data can be observed. For example there is a partial similarity between the fourth sequence from Table V.1 and the last sequence from Table V.2.

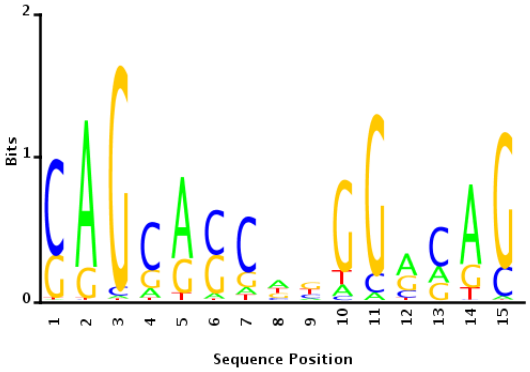
Figure V.1(c) shows the sequence logo resulted by running the data through ChIP-Munk [32] (A program for finding TFBSs from ChIP-seq data). We can observe very high similarity between the logo resulted by this study and the logo generated by



(a) NRSF logo from table V.1 for the NRSF motifs generated by our pipeline



(b) NRSF logo from TRANSFAC



(c) NRSF logo from running data on ChIPMunk [32]

Figure V.1: NRSF logos comparison.

ChIPMunk. Also, the information content for both our motif family and ChIPMunk motif family is comparable.

Moving to GABP TF, Table V.3 shows the top ranked TFBSs using shuffled train data as background. By looking at the first sequence in Table V.3, we observe that there is very high similarity between the motif consensus and the consensus reported by TRANSFAC. This is shown in V.2(a) and V.2(c).

In terms of motifs quality, binding sites reported using shuffled data as a background and TRANSFAC binding sites show consistent quality in terms of IC. It is

Table V.2: Top 5 ranked NRSF TFBSs based on cell line control data

Predicted model consensus	Score	P-value	Train coverage	Test coverage
GGGCAGRGGCG	1.84	1.47037e-119	0.890	0.874
GGGAGGCRGAG	1.81	1.65479e-116	0.877	0.858
KGGTGCTGARG	1.80	2.50872e-127	0.853	0.849
GAGGCKGRGGC	1.80	1.18639e-115	0.877	0.854
GCCYCMGCCTC	1.78	1.18639e-115	0.865	0.852

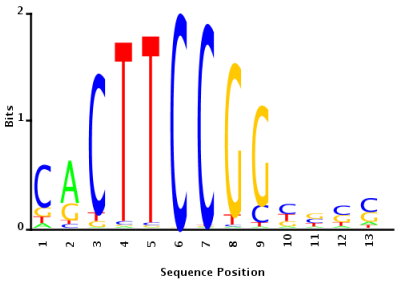
interesting to note that TRANSFAC GABP binding sites are derived from ChIP-seq data. Also, the length of both binding motifs is similar.

Table V.3: Top 5 ranked GABP TFBSs based on shuffled background

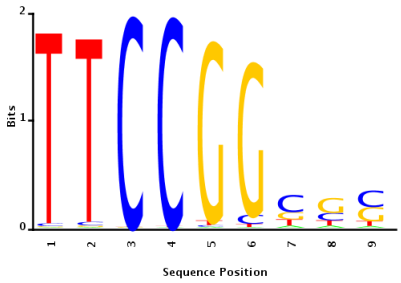
Predicted model consensus	Score	P-value	Train coverage	Test coverage
CACTTCCGGCSCC	1.989	2.15856e-120	0.958	0.94
ACTTCCG	1.988	7.95426e-105	0.976	0.976
CCGGAAGTGGC	1.970	3.02955e-126	0.93	0.94
SCGCCACTTCCGGCSCC	1.960	2.72814e-121	0.944	0.930
GGSGCCGGAAGTG	1.960	2.15856e-120	0.936	0.94

On the other hand, by looking to the other motifs in Table V.3, it is obvious that the consensus motif CACTTCCGGCSCC of the first motif family is contained within the consensus motif of the fourth motif family covering positions from 5 to the end of consensus sequence. The second, third, and fifth motifs contain also common subsequences. The motif ACTTCCG is present in the second motif and the reverse complement of the third and the fifth motif implying that the motif ACTTCCG is a strong candidate binding site for another TF that could bind together with GABP in the process of gene regulation in that cell. Also, it could be an alternative TFBS for GABP.

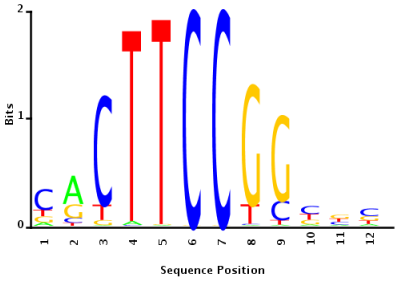
In the case when cell line control data is used as the background, the pipeline also identified the desired TFBS but only partially (see sequence logo in Figure V.2(b)).



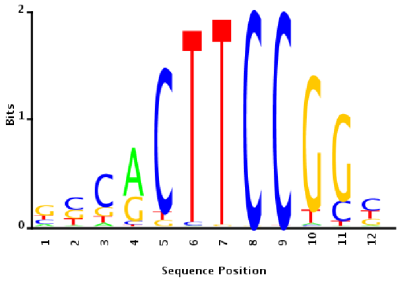
(a) GABP logo from table V.3 from motifs derived by our pipeline



(b) GABP logo from table V.4 from motifs derived by our pipeline



(c) GABP logo from TRANSFAC



(d) GABP logo from running data on ChIPMunk [32]

Figure V.2: GABP logos comparison.

The reported motif shows a high similarity to TRANSFAC motif from positions 4 to 12. The IC for the reported binding site is also comparable to the reference IC, since the TRANSFAC version of GABP binding sites is derived from a large number of 500 ChIP-seq identified TFBS data, so the discrepancy in the number of TFBSs used in our method that those of TRANSFAC is not that dramatic as it was for NRSF.

Table V.4: Top 5 ranked GABP TFBSs based on cell line control data

Predicted model consensus	Score	P-value	Train coverage	Test coverage
TTCCGGCGS	3.130	0	0.956	0.950
ACTTCCGGCSCC	3.100	0	0.958	0.941
GGMGCCGGAAGT	3.309	0	0.957	0.941
CGCCGGAAG	3.308	0	0.938	0.958
GCCACTTCCGGC	3.307	0	0.954	0.940

It is shown in Table V.4 that all p-values for the reported motif families are zero,

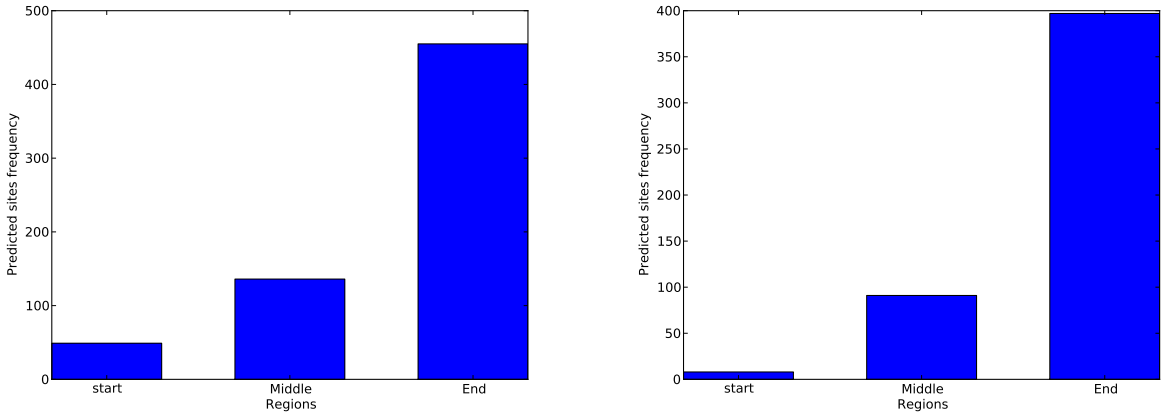
which implies that the only factor that affects the models ranking is the coverage of the model on the training and the testing data. Common motifs are also found in this experiment among the top 5 motif families. The motif TCCGGC is common between the first motif, the fifth motif, the third motif reverse complement and the fourth motif reverse complement.

Figure V.2(d) shows the sequence logo generated by the ChIPMunk program on GABP TF data. The logo shows similarity to the logo generated by our approach and the one from TRANSFAC, but the sequence is shifted to right by two nucleotides which makes it misses the first two nucleotides. It is however not possible to say which of the models is more correct.

By analyzing these results, the model selection based on the shuffled training data shows a better performance over the model selection based on the cell line experimental data. This conclusion is based on the two datasets that are used in this study, so it is not possible to generalize it. A superficial view of the situation should lead to the conclusion that the cell line experimental data is better than the artificially generated data. However, artificially shuffled data used as background results in more precision in the predicted motifs compared to cell line data, at least in the experiments we made.

One more type of analysis has been made, in which we have tried to check for a pattern in the distribution of the predicted TFBSs in both datasets. Figures V.3(a) and V.3(b) show the distribution for the predicted sites for NRSF and GABP TFs. To obtain these distributions the fragments (which are of different length) are split into 3 equal parts corresponding to the start, middle and end part of the fragment.

Figure V.2 shows the same type of distribution of the location within the fragment of the predicted best motif. One can observe that the right end side of the ChIP-seq peak is highly enriched with these motifs. This pattern could have been obtained because ChIP-seq profiling stage does not perform shifting of the peaks from both



(a) Distribution for NRSF predicted binding sites (b) Distribution for GABP predicted binding sites

Figure V.3: Distributions of the predicted binding sites

strands. However, this observation requires a more detailed analysis that is not possible due to time limitation of this study.

Chapter VI

Conclusion

In this study, a computational methodology was developed to identify TFBSs from ChIP-seq data. The method defines three major steps, which are ChIP-seq data processing, motif discovery and models selection. It showed high quality prediction of TFBSs based on comparison with the TRANSFAC reference binding sites. This study contributed the new motif family ranking strategy for ab-initio motif families identified from ChIP-seq fragments.

We showed that, through the use of our system, it is possible to predict other binding sites that may work in synergy with the target TF in gene regulation. Also, our method produces better accuracy results when artificially shuffled ChIP-seq fragments are used as the background, as compared to the situation when the cell line background data is used for the same purpose.

The pipeline system developed here may be further extended to point to the way how to construct a more efficient global scoring function for models selection and model ranking. Finally, implementing other types of TFBS models may lead to enhanced accuracy of the proposed method.

REFERENCES

[1] Droit A, Cheung C, and Gottardo R. rMAT–an R/Bioconductor package for analyzing ChIP-chip experiments. *Bioinformatics (Oxford, England)*, 26(5):678–9, March 2010.

[2] Sandelin A, Alkema W, Engstrm P, Wasserman WW, and Lenhard B. Jaspar: an openaccess database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(suppl 1):D91–D94, 2004.

[3] Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, and Sidow A. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods*, 5(9):829–834, 2008.

[4] Patterson AL. Ca direct method for the determination of the components of interatomic distances in crystals. *Zeitschrift fr Kristallographie*, 90:517, 1935.

[5] Boyle AP, Guinney J, Crawford GE, and Furey TS. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics (Oxford, England)*, 24(21):2537–8, November 2008.

[6] Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, and Jones SJ. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics (Oxford, England)*, 24(15):1729–30, August 2008.

[7] Marchand B, Bajic VB, and Kaushik DK. Highly Scalable Ab Initio Genomic Motif Identification. A paper submitted to SuperComputing'11 conference, May 2011.

[8] Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, and Young RA. Genome-wide location and function of dna binding proteins. *Science*, 290(5500):2306–2309, 2000.

[9] Yoav B and Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[10] Tuerk C and Gold L. Systematic evolution of ligands by exponential enrichment: Rna ligands to bacteriophage t4 dna polymerase. *Science*, 249(4968):505–510, 1990.

[11] Zang C, Schones DE, Zeng C, Cui K, Zhao K, and Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics (Oxford, England)*, 25(15):1952–8, August 2009.

[12] Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, and Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science (New York, N.Y.)*, 262(5131):208–14, October 1993.

[13] The ENCODE Project Consortium. The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640, 2004.

[14] Galas DJ and Schmitz A. Dnaase footprinting a simple method for the detection of protein-dna binding specificity. *Nucleic Acids Research*, 5(9):3157–3170, 1978.

[15] Gumucio DL, Heilstedt-Williamson H, Gray TA, Tarl SA, Shelton DA, Tagle DA, Slightom JL, Goodman M, and Collins FS. Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. *Mol. Cell. Biol.*, 12(11):4919–4929, 1992.

[16] Latchman DS. Transcription factors: An overview. *The International Journal of Biochemistry and Cell Biology*, 29(12):1305 – 1312, 1997.

[17] Huang E, Yang L, Chowdhary R, KAssim A, and Bajic VB. An algorithm for ab initio dna motif detection. *Information processing and living systems*, pages 611–615, 2006.

[18] Weisstein EW. Fisher's exact test. from mathworld. a wolfram web resource. http://mathworld.wolfram.com/fishersexacttest.html.

[19] Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, and Jones S. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, 4(8):651–657, 2007.

[20] Tuteja G, White P, Schug J, and Kaestner KH. Extracting transcription factor targets from ChIP-Seq data. *Nucleic acids research*, 37(17):e113, September 2009.

[21] Stormo GD. DNA binding sites: representation and discovery. *Bioinformatics (Oxford, England)*, 16(1):16–23, January 2000.

[22] Hertz GZ and Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics (Oxford, England)*, 15(7-8):563–77, 1999.

[23] Ji H, Jiang H, Ma W, Johnson DS, Myers RM, and Wong WH. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nature biotechnology*, 26(11):1293–300, November 2008.

[24] Li H and Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5):589–95, March 2010.

[25] Pearson H. Genetics: What is a gene? *Nature*, 441:398–401, 2006.

[26] Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, and Gerstein MB. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology*, 27(1):66–75, January 2009.

[27] Gralla JD. Activation and repression of e. coli promoters. *Current Opinion in Genetics and Development*, 6(5):526 – 530, 1996.

[28] Watson JD. *Molecular Biology of the gene*. Pearson, San Fransisco, CA, sixth edition edition, 2008.

[29] Song JS, Johnson WE, Zhu X, Zhang X, Li W, Manrai AK, Liu JS, Chen R, and Liu XS. Model-based analysis of two-color arrays (MA2C). *Genome biology*, 8(8):R178, January 2007.

[30] Ho JW, Bishop E, Karchenko PV, Ngre N, White KP, and Park PJ. ChIP-chip versus ChIP-seq: Lessons for experimental design and data analysis. *BMC genomics*, 12(1):134, February 2011.

[31] Connaghan-Jones KD, Moody AD, and Bain DL. Quantitative DNase footprint titration: a tool for analyzing the energetics of protein-DNA interactions. *Nature protocols*, 3(5):900–14, January 2008.

[32] Kulakovskiy, IV, Boeva VA, Favorov AV, and Makeev VJ. Deep and wide digging for binding motifs in chip-seq data. *Bioinformatics*, 26(20):2622–2623, 2010.

[33] Duret L and Bucher P. Searching for regulatory elements in human noncoding sequences. *Current Opinion in Structural Biology*, 7:399–406, 1997.

[34] O'Neill LP and Turner BM. Immunoprecipitation of chromatin. In Adhya S, editor, *RNA Polymerase and Associated Factors, Part B*, volume 274 of *Methods in Enzymology*, pages 189 – 197. Academic Press, 1996.

[35] Blanchette M and Tompa M. Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting. *Genome Research*, pages 739–748, 2002.

[36] Hu M, Yu J, Taylor JM, Chinnaiyan AM, and Qin ZS. On the detection and refinement of transcription factor binding sites using chip-seq data. *Nucleic Acids Research*, 38(7):2154–2167, 2010.

[37] Jiang M, Anderson J, Gillespie J, and Mayne M. ushuffle: A useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, 9(1):192, 2008.

[38] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, and Higgins DG. Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, 23(21):2947–8, November 2007.

[39] Das MK and Dai HK. A survey of dna motif finding algorithms. *BMC Bioinformatics*, 8(Suppl 7):S21, 2007.

[40] Bulyk ML. Computational prediction of transcription-factor binding site locations. *Genome biology*, 5(1):201, January 2003.

[41] Mendes ND, Casimiro AC, Santos PM, S-Correia I, Oliveira AL, and Freitas AT. Musa: a parameter free algorithm for the identification of biologically significant motifs. *Bioinformatics*, 22(24):2996–3002, 2006.

[42] Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10):669–80, October 2009.

[43] DeHaseth PL, Gross CA, and Record MT Jr. Burgess RR. Measurement of binding constants for protein-dna interactions by dna-cellulose chromatography. *Biochemistry*, 16:4777–4783, 1977.

[44] Jothi R, Cuddapah S, Barski A, Cui K, and Zhao K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic acids research*, 36(16):5221–31, September 2008.

[45] SIMES RJ. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[46] Georgiev S, Boyle AP, Jayasurya K, Ding X, Mukherjee S, and Ohler U. Evidence-ranked motif identification. *Genome Biology*, 11(2):R19, 2010.

[47] Neph S and Tompa M. Microfootprinter: a tool for phylogenetic footprinting in prokaryotic genomes. *Nucleic Acids Research*, 34(suppl 2):W366–W368, 1 July 2006.

[48] Pepke S, Wold B, and Mortazavi A. computation for chIP-seq and rNA-seq studies. *Signals*, 6(11), 2009.

[49] Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, October 1990.

[50] Kouzarides T. Chromatin modifications and their function. *Cell*, 128(4):693 – 705, 2007.

[51] Bailey TL and Elkan C. The value of prior knowledge in discovering motifs with MEME. *Proceedings of International Conference on Intelligent Systems for Molecular Biology*, 3:21–29, 1995.

[52] Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, and Bernstein BE. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, July 2007.

[53] Matys V, Fricke E, Geffers R, Gssling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Mnch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, and Wingender E. Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378, 2003.

[54] Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, and Liu XS. Model-based analysis of tiling-arrays for ChIP-chip. *Sciences-New York*, 2006.

[55] Kent WJ. BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12(4):656–664, March 2002.

[56] Qi Y, Rolfe A, MacIsaac KD, Gerber GK, Pokholok D, Zeitlinger J, Danford T, Dowell RD, Fraenkel E, Jaakkola TS, Young RA, and Gifford DK. High-resolution computational models of genome binding events. *Nature biotechnology*, 24(8):963–70, August 2006.

[57] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, and Liu XS. Model-based analysis of ChIP-Seq (MACS). *Genome biology*, 9(9):R137, January 2008.